

#4

35.C14035

PATENT APPLICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:)
TAKAFUMI MIZUNO) : Examiner: NYA
Application No.: 09/449,706) : Group Art Unit: 2776
Filed: November 24, 1999) :
For: DOCUMENT TYPE DEFINITION)
GENERATING METHOD AND :
APPARATUS, AND STORAGE)
MEDIUM FOR STORING :
PROGRAM) February 4, 2000

Assistant Commissioner for Patents
Washington, D.C. 20231

CLAIM TO PRIORITY

Sir:

Applicant hereby claims priority under the
International Convention and all rights to which he is entitled
under 35 U.S.C. § 119 based upon the following Japanese Priority
Application:

10-336378 filed November 26, 1998

A certified copy of the priority document is
enclosed.

Applicant's undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our address given below.

Respectfully submitted,


Attorney for Applicant

Registration No. 25,823

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, New York 10112-3801
Facsimile: (212) 218-2200

NY_MAIN 59171 v 1

Cfo 14035 (EP) US,
/m

09/449.706



日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
いる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
with this Office.

出 願 年 月 日
Date of Application:

1998年11月26日

出 願 番 号
Application Number:

平成10年特許願第336278号

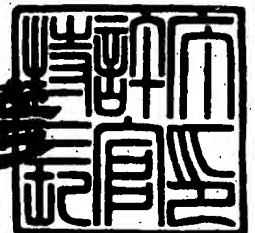
願 人
Applicant (s):

キヤノン株式会社

1999年12月17日

特許庁長官
Commissioner,
Patent Office

近 藤 隆 彦



【書類名】 特許願

【整理番号】 3798113

【提出日】 平成10年11月26日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/27

【発明の名称】 文書型定義生成方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体

【請求項の数】 24

【発明者】

【住所又は居所】 東京都大田区下丸子3丁目30番2号 キヤノン株式会社内

【氏名】 水野 貴史

【特許出願人】

【識別番号】 000001007

【氏名又は名称】 キヤノン株式会社

【代理人】

【識別番号】 100077481

【弁理士】

【氏名又は名称】 谷 義一

【選任した代理人】

【識別番号】 100088915

【弁理士】

【氏名又は名称】 阿部 和夫

【手数料の表示】

【予納台帳番号】 013424

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9703598

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書型定義生成方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体

【特許請求の範囲】

【請求項 1】 複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（1 行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断ステップと、

前記タグの意味的構造を判断する意味的構造判断ステップとを有し、

前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする文書型定義生成方法。

【請求項 2】 前記タグの物理的構造の判断ステップにおいて、タグの位置関係を認識することを特徴とする請求項 1 に記載の文書型定義生成方法。

【請求項 3】 前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや 1 行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理も含むことを特徴とする請求項 2 に記載の文書型定義生成方法。

【請求項 4】 前記タグの意味的構造の判断ステップにおいて、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出することを特徴とする請求項 1 ないし 3 のいずれかに記載の文書型定義生成方法。

【請求項 5】 前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除ステップを有することを特徴とする請求項 1 ないし 4 のいずれかに記載の文書型定義生成方法。

【請求項 6】 前記冗長性排除ステップは、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行うことを特徴

とする請求項 5 に記載の文書型定義生成方法。

【請求項 7】 前記冗長性排除ステップは、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いることを特徴とする請求項 6 に記載の文書型定義生成方法。

【請求項 8】 同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成ステップを有することを特徴とする請求項 1 ないし 7 のいずれかに記載の文書型定義生成方法。

【請求項 9】 複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれていた前後の行（1 行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断手段と、

前記タグの意味的構造を判断する意味的構造判断手段とを有し、

前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする文書処理装置。

【請求項 10】 前記タグの物理的構造の判断手段は、タグの位置関係を認識することを特徴とする請求項 9 に記載の文書処理装置。

【請求項 11】 前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや 1 行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理も含むことを特徴とする請求項 10 に記載の文書処理装置。

【請求項 12】 前記タグの意味的構造の判断手段は、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出することを特徴とする請求項 9 ないし 11 のいずれかに記載の文書処理装置。

【請求項 13】 前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは

同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除手段を有することを特徴とする請求項9ないし12のいずれかに記載の文書処理装置。

【請求項14】 前記冗長性排除手段は、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行うことを特徴とする請求項13に記載の文書処理装置。

【請求項15】 前記冗長性排除手段は、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いることを特徴とする請求項14記載の文書処理装置。

【請求項16】 同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成手段を有することを特徴とする請求項9ないし15のいずれかに記載の文書処理装置。

【請求項17】 コンピュータによって実行される文書型定義生成用プログラムを記録した記録媒体であって、該プログラムはコンピュータに対し、

複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（1行以上）の字下げ、空白行という物理的構造を判断させ、

前記タグの意味的構造を判断させ、

前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成させることを特徴とする文書型定義生成用プログラムを記録した記録媒体。

【請求項18】 前記プログラムはコンピュータに対し、

前記タグの物理的構造の判断において、タグの位置関係を認識させることを特徴とする請求項17記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項19】 前記プログラムはコンピュータに対し、

前記タグの位置関係の認識において、字下げを行うことによって引用を表す引

用中に行われる字下げや 1 行あるいは複数行を常に飛ばしている場合のような物理的構造検出によって無意味な構造を取り除く処理行わせることを特徴とする請求項 18 に記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項 20】 前記プログラムはコンピュータに対し、

前記タグの意味的構造の判断において、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出させることを特徴とする請求項 17 ないし 19 のいずれかに記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項 21】 前記プログラムはコンピュータに対し、

前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求めさせ、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除させることを特徴とする請求項 17 ないし 20 のいずれかに記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項 22】 前記プログラムはコンピュータに対し、

前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行わせることを特徴とする請求項 21 に記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項 23】 前記プログラムはコンピュータに対し、

タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いさせることを特徴とする請求項 22 に記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項 24】 前記プログラムはコンピュータに対し、

同一名称の開始タグと終了タグ間に存在する文の語句解析を行わせ、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行わせることを特徴とする請求項 17 ないし 23 のいずれかに記載の文書型定義生成用プログラムを記録した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、パーソナルコンピュータやワードプロセッサ等で実行される電子化文書の処理技術に関し、特に構造化文書の文書型定義を生成する方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体に関する。

【0002】

【従来の技術】

近年、パーソナルコンピュータやワードプロセッサ等で作成された電子化された文書が溢れかえっている。このような電子化文書に対して、統一的に扱え、文書に意味情報を付与する構造化文書の導入が進められている。しかしながら、この構造化文書を記述するには、予め決められた規則（文書型定義）に従わなければならないことが多い。

【0003】

しかし、従来では、タグの置かれている物理的観点、意味的観点から文書型定義を生成する方法はなかった。

【0004】

【発明が解決しようとする課題】

文書型定義を予め用意しておかなくても、構造化文書を記述できるが、文書の記述を行うユーザ各々が勝手にタグを使ってしまうと、タグに付与された意味情報も正しく扱うことができなくなってしまう。また、逆に同一の種類と考えられる文書に対して複数の文書型定義が用意されている場合には、タグに対する冗長性が存在してしまう可能性がある。

【0005】

本発明の目的は、上記のような課題を解決すべく、タグに与えられた意味情報を正しく扱うことができ、タグに対する冗長性を排除した文書型定義を生成することが可能な文書型定義生成方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体を提供することにある。

【0006】

【課題を解決するための手段】

上記目的を達成するため、請求項1の文書型定義生成方法の発明は、複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれてある前後の行（1行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断ステップと、前記タグの意味的構造を判断する意味的構造判断ステップとを有し、前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする。

【0007】

ここで、好ましくは、前記タグの物理的構造の判断ステップにおいて、タグの位置関係を認識する。

【0008】

また、好ましくは、前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや1行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理も含む。

【0009】

また、好ましくは、前記タグの意味的構造の判断ステップにおいて、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出する。

【0010】

また、好ましくは、前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除ステップを有する。

【0011】

また、好ましくは、前記冗長性排除ステップは、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行う。

【0012】

また、好ましくは、前記冗長性排除ステップは、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いる。

【0013】

また、好ましくは、同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成ステップを有する。

【0014】

上記目的を達成するため、請求項9の文書処理装置の発明は、複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（1行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断手段と、前記タグの意味的構造を判断する意味的構造判断手段とを有し、前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする。

【0015】

ここで、好ましくは、前記タグの物理的構造の判断手段は、タグの位置関係を認識する。

【0016】

また、好ましくは、前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや1行あるいは複数行を常に飛ばしている場合のような物理的構造検出によって無意味な構造を取り除く処理も含む。

【0017】

また、好ましくは、前記タグの意味的構造の判断手段は、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出する。

【0018】

また、好ましくは、前記物理的構造判断と前記意味的構造判断の結果に元づい

て、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除手段を有する。

【0019】

また、好ましくは、前記冗長性排除手段は、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行う。

【0020】

また、好ましくは、前記冗長性排除手段は、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いる。

【0021】

また、好ましくは、同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成手段を有する。

【0022】

上記目的を達成するため、請求項17の記録媒体の発明は、コンピュータによって実行される文書型定義生成用プログラムを記録した記録媒体であって、該プログラムはコンピュータに対し、複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（1行以上）の字下げ、空白行という物理的構造を判断させ、前記タグの意味的構造を判断させ、前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成させることを特徴とする。

【0023】

また、好ましくは、前記プログラムはコンピュータに対し、前記タグの物理的構造の判断において、タグの位置関係を認識させる。

【0024】

また、好ましくは、前記プログラムはコンピュータに対し、前記タグの位置関

係の認識において、字下げを行うことによって引用を表す引用中に行われる字下げや1行あるいは複数行を常に飛ばしている場合のような物理的構造検出によって無意味な構造を取り除く処理行わせる。

【0025】

また、好ましくは、前記プログラムはコンピュータに対し、前記タグの意味的構造の判断において、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出させる。

【0026】

また、好ましくは、前記プログラムはコンピュータに対し、前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求めさせ、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除させる。

【0027】

また、好ましくは、前記プログラムはコンピュータに対し、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行わせる。

【0028】

また、好ましくは、前記プログラムはコンピュータに対し、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いさせる。

【0029】

また、好ましくは、前記プログラムはコンピュータに対し、同一名称の開始タグと終了タグ間に存在する文の語句解析を行わせ、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行わせる。

【0030】

本発明は、上記構成により、タグの置かれている前後の行の字下げ、空白行という物理的構造を判断し、タグの意味的構造を判断して、タグの位置関係を認識することにより文書の型定義を行うようにしているので、タグに与えられた意味

情報を正しく扱うことが可能となる。

【0031】

更に、本発明では、タグに対する冗長性の排除、及び、同一名称で異なる意味のタグの名称変更を行った文書型定義を生成することが可能となる。

【0032】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態を詳細に説明する。

【0033】

(第1の実施の形態)

本発明に係る文書型定義生成方法を実現する一実施形態の処理手順を図1のフローチャートに示す。なお、以下に述べる手順は、プログラムコード形態でフロッピディスク(FD)、ハードディスク、光ディスク、光磁気ディスク、CD-ROM等の記録媒体に記録されることができ、CPUで読み出されて実行されるものとする。また、この文書型定義生成方法は一般的なハード構成のパーソナルコンピュータやワードプロセッサ等で実行可能であるので、この方法を実行するハードウェアの具体的なシステム構成の説明は省略する。

【0034】

まず、ステップ101において構造化文書の入力が行われる。ここで与える構造化文書の一例を図2の(a)に示す。この構造化文書の中から、次のステップ102において各々のタグ位置が検出され、最初に“<Title>”から順番にタグ番号が振られる。

【0035】

続いて、ステップ103において、上記文書における物理的構造の検出が行われる。例えば、図2の(b)において、与えた段落を表すタグ“<Para>”において模式的に表現されているように、字下げで始まる文章群を段落と見なすという特徴の検出を行う。このような物理的構造を検出するための処理手順一例を図3のフローチャートに示す。

【0036】

ステップ31で文書中の字下げが行われている行を見つけ出し、次のステップ

32でその行に続く文章群を検出する。このときに字下げが行われている行から次の字下げが行われている行、あるいは空白行の直前の行までを文章群とすることができる。また、字下げを行うことによって引用を表す引用中に行われている字下げ（2重の字下げ）や、一行あるいはそれ以上に行を常に飛ばして記述されている場合には、ステップ32の中で、文書全体のパターンから、物理的構造検出にとって無意味な構造を取り除いて処理を行う。

【0037】

図1のステップ104では、入力された構造化文書が持つ意味的構造を検出する。一例として、上記の図2の(a)において、“<Section>”タグは、先頭に「1.」、「2.」、「3.」といった形式をしている。ここで意味的に、“<Section>”タグは、「数字.」からなると推察できる。この意味的構造を検出する処理手順の一例を図4のフローチャートに示す。

【0038】

まず、ステップ41において、文書中の全ての単語、記号について意味情報DB（データベース）43と通信をして、文書中の語間の繋がりや単語・記号の種類を与える。この結果を元に、次のステップ42において、各タグにおいて見られる意味的な構造を検出する。

【0039】

次に、図1のステップ105において、処理すべきタグの開始を最初に出現したタグとし、ステップ106でタグの処理が全て終了したか否かを判断する。

【0040】

タグの処理が全て終了していない場合は、ステップ107に移行して、現在処理しているタグと、上記ステップ103，ステップ104において検出された物理的構造・意味的構造の情報との統合を行う。ここで、統合とは、現在、処理しているタグに関わる行に物理的な、意味的な特徴が存在するならば、タグとその情報を結びつけることをいう。次のステップ108では、次に出現したタグに処理を移し、上記ステップ106に戻る。

【0041】

ステップ106においてタグの処理が全て終了したと判断した場合は、ステッ

ブ 109 に移行し、異なった名称のタグの間において類似度を求め、その類似度が予め設定したあるしきい値以上の場合にはそれらタグは互いに同一のタグと見なして、一方のタグを生成する文書型定義に出現しないようにする。この類似度を求めて同じ内容を持つタブであるか否かを判定する処理手順の一例を図 5 のフローチャートに示す。

【0042】

まず、ステップ 51 では、異なる名称をもつタグ A, B の類似度を算出する。この算出方法は、物理的構造が一致する場合は 1、物理的構造が完全に一致しないが一部が一致する場合は、その一致した割合に対応する 1 未満の値にし、意味的構造についても同様の考えをあてはめ、その和を 2 で割ったものを類似度 d_{AB} とする。

【0043】

ステップ 51 で求めた上記の類似度 d_{AB} に対して、次のステップ 52 において、予め与えたしきい値 δ との比較を行う。その類似度 d_{AB} が δ 未満のときは、ステップ 54 に飛んで、次の組み合わせを試みる。

【0044】

その類似度 d_{AB} が δ より大きいときは、ステップ 53 に移行して、最終的に文書型定義を生成するリストからタグ B を削除し、冗長性の排除を行う。また、フローチャートとして示さないが、同一名称を持つ異なる意味のタグの名称変更も図 1 のステップ 109 中で行う。

【0045】

ステップ 53 の処理が終了したらステップ 54 へ進み、全てのタグの組み合わせを試したか否かを判断し、全てのタグの組み合わせを試していない場合はステップ 51 に戻り、上記と同様に、タグ A a , A b 間での類似度を求めて、その類似度の値 d_{AaAb} がしきい値 δ 未満のときは、タグ A b の名称の変更を行う。全てのタグの組み合わせを試した場合は本サブルーチンの処理を終了して図 1 のメインルーチンに戻る。

【0046】

図 1 のステップ 110 では、同一名称の開始タグと終了タグ間の文の語句につ

いての解析を行い、タグ中に含まれるべき情報を得る。この解析結果を用いて、次のステップ 111 において、文書型定義の生成を行う。

【0047】

(第2の実施の形態)

上述した本発明の第1の実施形態においては、文書中の物理の構造、意味的構造を本文（タグ以外の文）に対して行ってきたが、本発明はその限りではない。

【0048】

例えば、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を利用して類似度の尺度として用いてもよい。

【0049】

(他の実施形態)

なお、本発明は、複数の機器（例えば、ホストコンピュータ、インターフェース機器、リーダ、プリンタなど）から構成されるコンピュータシステムに適用しても、1つの機器からなる装置（例えば、ワードプロセッサ、複写機、ファクシミリ装置など）に適用してもよい。

【0050】

また、本発明の目的は、前述した実施の形態の機能を実現するソフトウェアのプログラムコードを記録した記録媒体（記憶媒体）を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはCPUやMPU）が記録媒体に格納されたプログラムコードを読み出し、実行することによっても、達成されることは言うまでもない。

【0051】

この場合、記録媒体から読み出されたプログラムコード自体が前述した実施の形態の機能を実現することになり、そのプログラムコードを記録した記録媒体は本発明を構成することになる。

【0052】

そのプログラムコードを記録し、またテーブル等の変数データを記録する記録媒体としては、例えばフロッピディスク（FD）、ハードディスク、光ディスク

、光磁気ディスク、CD-ROM、CD-R、磁気テープ、不揮発性のメモ리카ード（ICメモ리카ード）、ROMなどを用いことができる。

【0053】

また、コンピュータが読み出したプログラムコードを実行することにより、前述の実施の形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づいて、コンピュータ上で稼動しているOS（オペレーティングシステム）などが実際の処理の一部または全部を行ない、その処理によって前述した実施の形態の機能が実現される場合も含まれることは言うまでもない。

【0054】

【発明の効果】

以上説明したように、本発明によれば、タグの置かれている前後の行の字下げ、空白行という物理的構造を判断し、タグの意味的構造を判断して、タグの位置関係を認識することにより文書の型定義を行うようにしているので、タグに与えられた意味情報を正しく扱うことが可能となる。

【0055】

更に、本発明によれば、タグに対する冗長性の排除、及び、同一名称で異なる意味のタグの名称変更を行った文書型定義を生成することが可能となる。

【図面の簡単な説明】

【図1】

本発明に係る文書型定義生成方法を実現する一実施形態の処理手順を示すフローチャートである。

【図2】

本システムに与える構造化文書データの一例を示す説明図である。

【図3】

図1のステップ103における物理的構造解析の処理手順の一例を示すフローチャートである。

【図4】

図1のステップ104における意味的構造解析の処理手順の一例を示すフローチャートである。

【図 5】

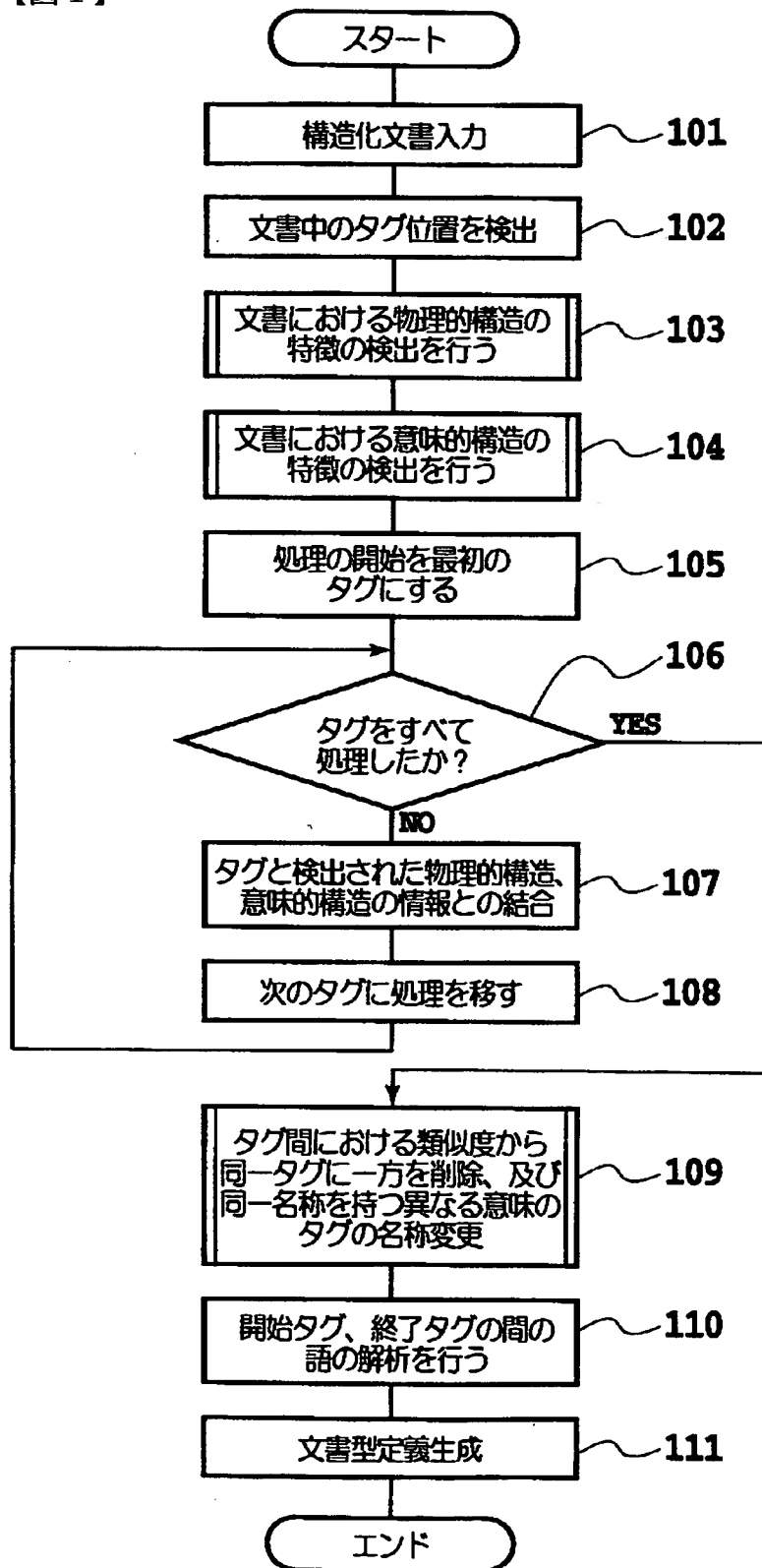
図 1 のステップ 109 におけるタグ間の類似度からタグの冗長性排除をする処理手順の一例を示すフローチャートである。

【符号の説明】

43 意味情報 DB (データベース)

【書類名】 図面

【図 1】



【図 2】

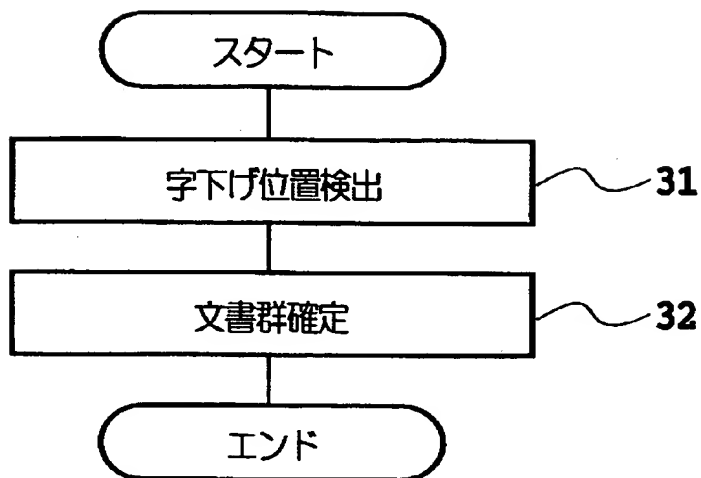
(a)

```
<Title>テレビの使用説明書</Title>
<Date> 1998.2.1</Date>
<Author>山田太郎</Author>
<Body>
  <Section>1.コンセントを差す</Section>
  <Section>2.電源を入れる </Section>
  <Section>3.チャンネルを合わせる</Section>
</Body>
```

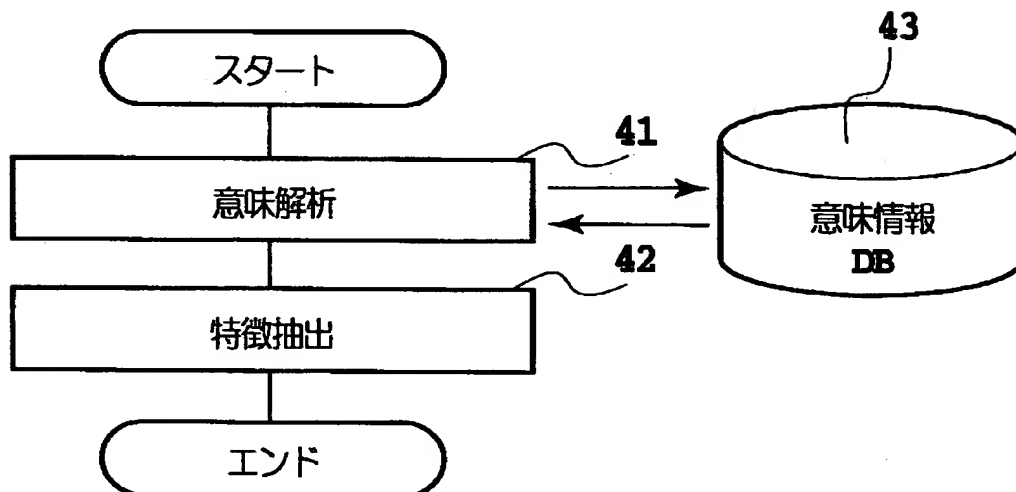
(b)

```
<Para>
  _____
  _____
  _____
</Para>
```

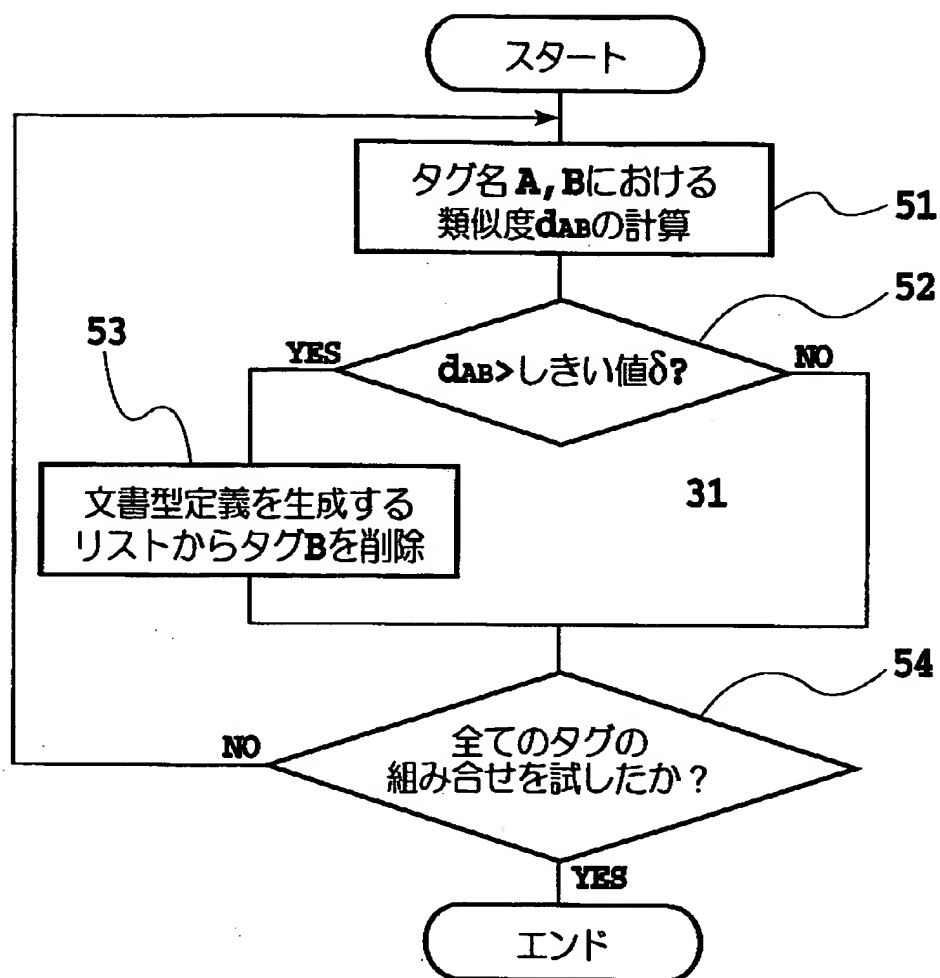
【図 3】



【図 4】



【図 5】



【書類名】 要約書

【要約】

【課題】 タグに与えられた意味情報を正しく扱うことができ、タグに対する冗長性を排除した文書型定義を生成する。

【解決手段】 構造化文書の文書構造を表現するために文書中に埋め込まれたタグのおかれている前後の行（1行以上）の字下げ、空白行という物理的構造を判断するステップ103と、タグの意味的構造を判断するステップ104と、異なった名称のタグ間において類似度を求め、類似度があるしきい値以上の場合は、同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除するステップ109と、同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、タグ中に含まれるべき情報を得て、文書型定義の生成を行うステップ110、111とを有する。

【選択図】 図1

【書類名】 職権訂正データ
【訂正書類】 特許願

<認定情報・付加情報>

【特許出願人】
【識別番号】 000001007
【住所又は居所】 東京都大田区下丸子3丁目30番2号
【氏名又は名称】 キヤノン株式会社
【代理人】 申請人
【識別番号】 100077481
【住所又は居所】 東京都港区赤坂2丁目6番20号 谷・阿部特許事
務所
【氏名又は名称】 谷 義一
【選任した代理人】
【識別番号】 100088915
【住所又は居所】 東京都港区赤坂2丁目6番20号 谷・阿部特許事
務所
【氏名又は名称】 阿部 和夫

出 願 人 履 歴 情 報

識別番号

[000001007]

1. 変更年月日	1990年 8月30日
[変更理由]	新規登録
住 所	東京都大田区下丸子3丁目30番2号
氏 名	キャノン株式会社